Bài 1 - Trực quan hóa dữ liệu (phần 1) (Data visualization)

Thống kê máy tính và ứng dụng (ĐTTX)

Vũ Quốc Hoàng
FIT - HCMUS

Ngày 22 tháng 12 năm 2021

Nội dung

1. Giới thiệu

2. Các công cụ và thư viện Python hỗ trợ

3. Các biểu đồ thông dụng

4. Các khái niệm cơ bản

Giới thiệu

- **Trực quan hóa dữ liệu** (Data visualization) là các kĩ thuật mô tả, biểu diễn, trình bày dữ liệu bằng hình ảnh, đồ họa.
- Gần gũi với **Trực quan hóa thông tin** (Information visualization).
- Là một bộ phận của **Thống kê mô tả** (Descriptive Statistics) hay **Phân tích dữ liệu thăm dò** (Exploratory data analysis - EDA).
- Là một lĩnh vực kết hợp: Nghệ thuật + Khoa học + Kỹ thuật + Công nghệ.

Các công cụ và thư viện Python hỗ trợ

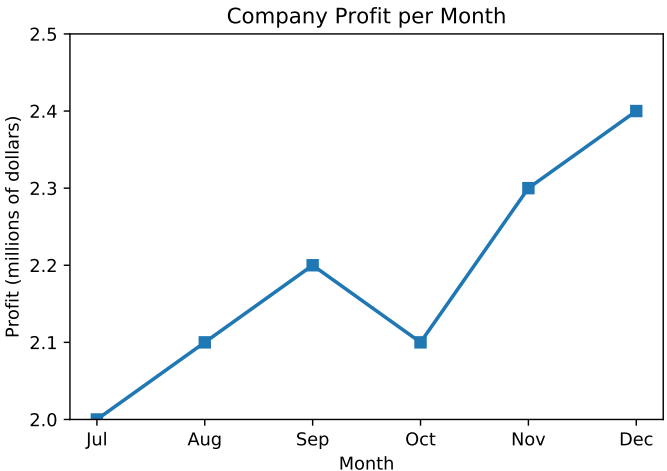
- Python (https://www.python.org/)
- Anaconda (https://www.anaconda.com/)
- Jupyter Notebook, Jupyter Lab (https://jupyter.org/)
- NumPy (https://numpy.org/)
- pandas (https://pandas.pydata.org/)
- Matplotlib (https://matplotlib.org/)
- seaborn (https://seaborn.pydata.org/)

Các biểu đồ thông dụng

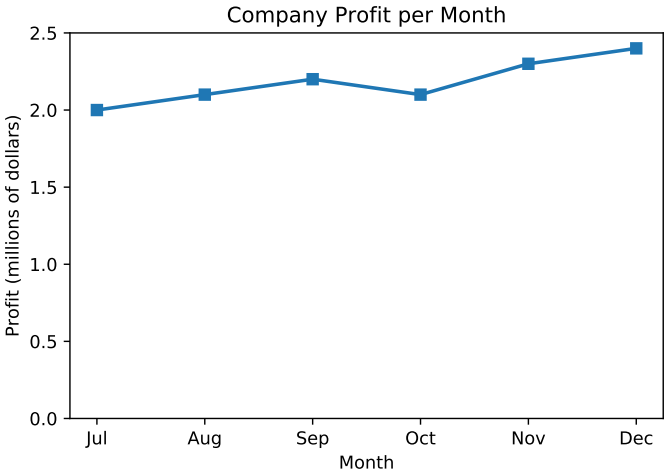
Phần này tìm hiểu các biểu đồ thông dụng nhất

- **Biểu đồ đường** (line chart): `https://en.wikipedia.org/wiki/Line_chart`
- **Biểu đồ quạt** (pie chart): `https://en.wikipedia.org/wiki/Pie_chart`
- **Biểu đồ thanh** (bar chart): `https://en.wikipedia.org/wiki/Bar_chart`

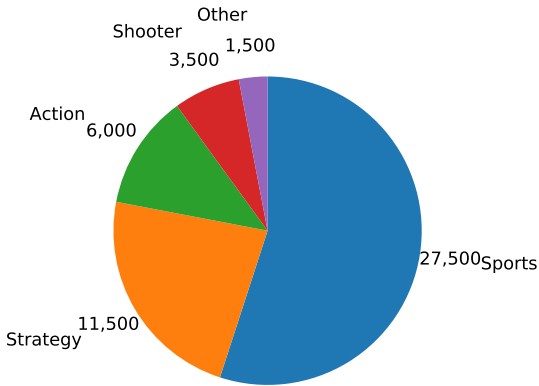
Biểu đồ đường (line chart)



Biểu đồ đường (tt)

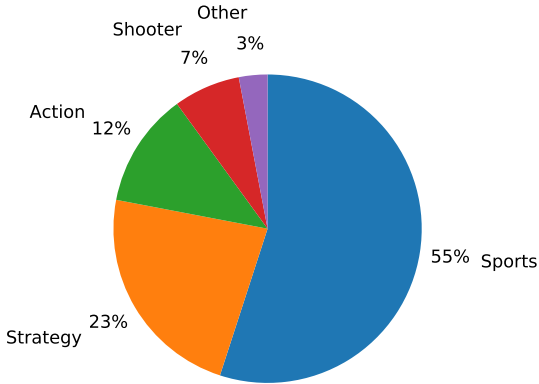


Biểu đồ quạt (pie chart)



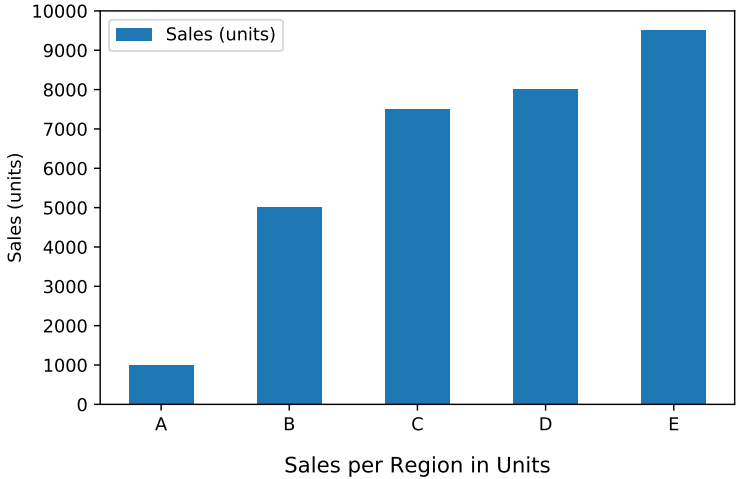
Units Sold per Genre

Biểu đồ quạt (tt)

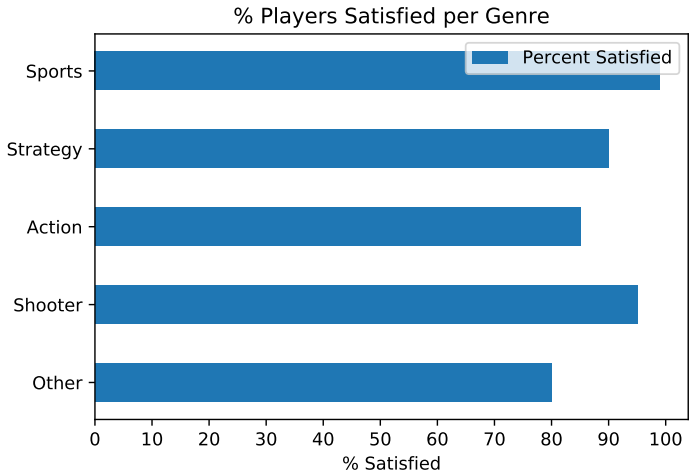


Units Sold per Genre

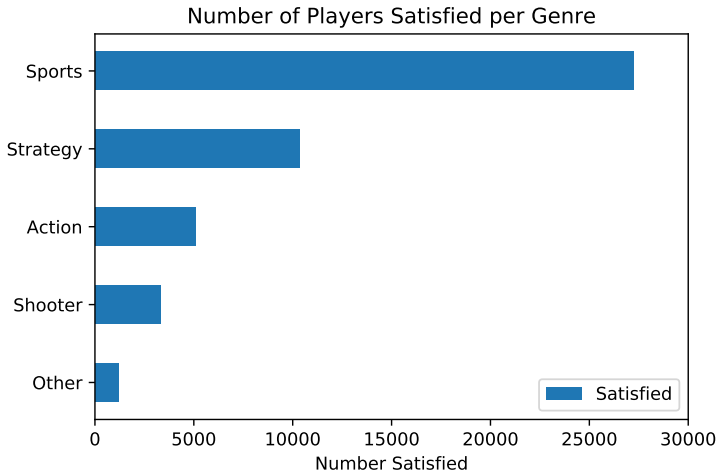
Biểu đồ thanh đứng (vertical bar chart)



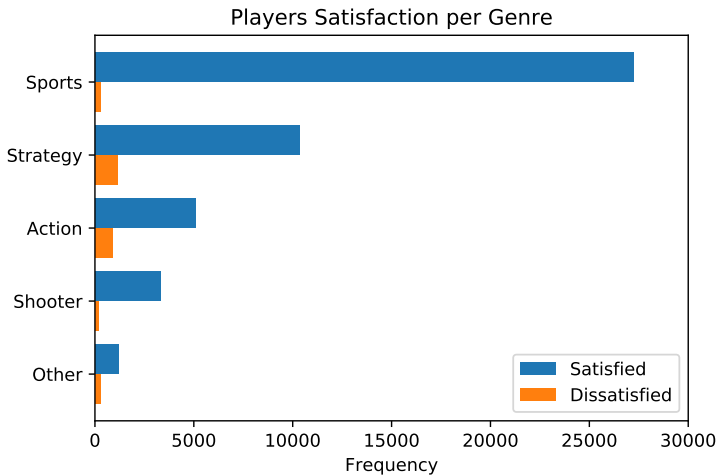
Biểu đồ thanh ngang (horizontal bar chart)



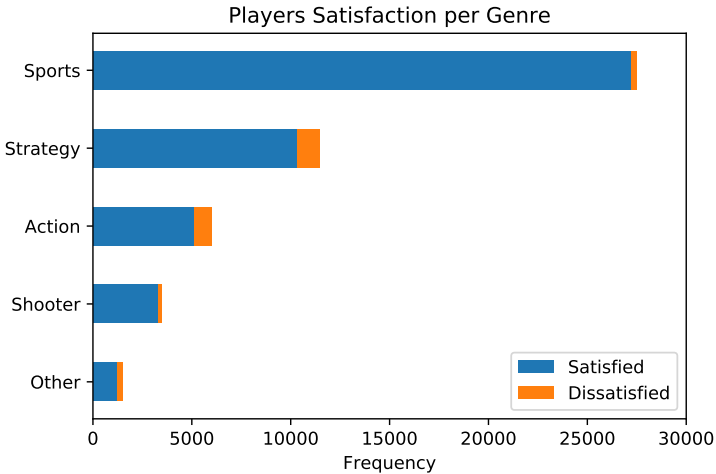
Biểu đồ thanh ngang (tt)



Biểu đồ thanh gom nhóm (grouped bar chart, split-category bar chart)



Biểu đồ thanh xếp chồng (stacked bar chart, segmented bar char)



Các khái niệm cơ bản

- **Dữ liệu** (data) là các con số, sự kiện, thông tin được cân đo, đong đếm, ghi nhận, quan sát từ các sự việc, hiện tượng muốn phân tích, nghiên cứu.
- **Mẫu dữ liệu** (data sample, data set, data table, database) là tập tất cả các dữ liệu liên quan thu thập được.
- Một **thống kê** (statistic, sample statistic) là một đại lượng (thường là một hay một vài con số) được tính toán từ mẫu dữ liệu.
- **Thống kê** (Statistics) là lĩnh vực khoa học liên quan đến việc thu thập, tổ chức, phân tích, giải thích và trình bày dữ liệu. Hoạt động thống kê gồm 3 bước chính
  - Thu thập dữ liệu (thiết kế thí nghiệm, thực hiện khảo sát, lấy từ nguồn sẵn có, ...)
  - Phân tích dữ liệu (dùng các kĩ thuật thống kê mô tả và thống kê suy diễn)
  - Rút ra kết luận (hiểu, dự đoán, ra quyết định, ...)
- Thống kê là cốt lõi của ngành **Khoa học dữ liệu** (Data science).

Tần số và tỉ lệ

- **Tần số** (frequency) là số lượng **đối tượng** (object, item) của một **nhóm** (group, category) nào đó.
- **Tỉ lệ** (proportion, percentage) hay **tần suất** (relative frequency) là tỉ lệ số lượng đối tượng của nhóm trên tổng số lượng.
- Ví dụ, bảng dữ liệu doanh số bán hàng (đơn vị cái)

Genre	Sports	Strategy	Action	Shooter	Other	**Total**
Units sold	27,500	11,500	6,000	3,500	1,500	**5,0000**
Percentage	55%	23%	12%	7%	3%	**100%**

Tài liệu tham khảo

Chapter 1. Dawn Griffiths. *Head First Statistics - A Brain-Friendly Guide*. O'Reilly Media, 2009.